Turn off the Radio and Call Again: How Acoustic Clues can Improve Dialogue Management

Christian Lieske
École Polytechnique Fédéderale Lausanne (EPFL)
e-mail : lieske@di.epfl.ch

1 Introduction

Traditionally, the most important input to the dialogue management component of a natural language system are semantic representations (e.g., formulae of first-order predicate calculus). Only recently, other kinds of information (e.g., phonetic transcriptions of unknown words) has been used. There is, however, room for the utilization of additional knowledge sources. In what follows, we first explicate how information from the accoustic level (e.g., the presence of certain kinds of background noise from a radio) enables better system performance. Our argumentation is that this is possible since the dialogue manager can use acoustic clues, on the one hand, to establish better recognition conditions, and on the other hand to generate more co-operative interactions. After this argumentation, we indicate how the acoustic cues for dialogue management components can be generated.

2 State-of-the-art in Dialogue Management

Today, many dialogue management components base their operations on a combination of dialogue history, a rule-based or statistical dialogue model, rule-based or statistical dialogue act idenfication rules, and semantic representations of user utterances. For spoken language input, this repertoire of information ensues interactions like the (unsuccessful) dialogue in Figure 1 (from a train timetable enquiry system).[1]

What happens in a dialogue like this is the following: The dialogue manager tries to verify a value that is needed for a database query. This attempt fails several times, since the incoming semantic representations are inconsistent due to recognition errors. After an arbitrary number of trials (two in the example), the dialogue manager surrenders.

More advanced systems exploit additional information to generate more intelligent feedback for the user. The Verbmobil device which translates spoken utterances between German and English as well as between Japanese and English (see (Kay et al., 1994)), e.g. exploits representations of unknown words and information about phonetically similar words to generate clarification dialogues along the lines of Figure 2 (cf. (Maier et al., 1997)).

Most of the ideas presented here were developed during work at Forum Technology, Malvern. I am indebted to all of my colleagues there, especicially Jörg Überla and David Bijl for providing valuable discussions. I also owe thanks to Afzal Ballim, and Yoshiki Mori for their comments on drafts of this poster.

[1] Apologies for the fact that all examples are related to German.

user	Ich möchte gerne um zwei Uhr nach Hamburg fahren.
user	*I would like to travel to Hamburg at two o'clock.*
system	Sie möchten nach Homburg fahren?
system	*You would like to go to Homburg?*
user	Nein, nicht nach Homburg, sondern nach Hamburg.
user	*No, not to Homburg but to Hamburg.*
system	Sie möchten nach Offenburg fahren?
system	*You would like to go to Offenburg?*
user	Nein, ich möchte auch nicht nach Offenburg, sondern nach Hamburg.
user	*No, I also don't want to go to Offenburg but to Hamburg.*
system	Es scheint ein Kommunikationsproblem zu geben. Bitte versuchen Sie es später noch einmal.
system	*We seem to have a communications problem. Please try again later.*

Figure 1: Communication failure

user	Mein Name ist Schwartz. Ich möchte mit Ihnen einen Termin im Juli ausmachen.
user	*My name is Schwartz. I would like to arrange an appointment in July.*
system	Ist Schwartz ein Eigenname?
system	*Is Schwartz a proper name?*
user	Ja.
user	*Yes.*
system	Sagten Sie 'Juli' oder 'Juni'?
system	*Did you say 'July' or 'June'?*
user	Juli.
user	*July.*

Figure 2: Intelligent dialogue

3 Dialogue Management with Acoustic Clues

Still missing even from the more sophisticated systems like Verbmobil is a flow of information all the way from the acoustic level up to the dialogue management component.[2] This kind of coupling will enable more successful systems for two reasons: improved speech recognition and more informative responses to the user.

Improvement in speech recognition will stem from the dialogue manager acting as a kind of mediator between the speech recognizer and the user. In case of bad recognition rates (speech recognizers already deliver confidence scores), the dialogue manager could ask for acoustic clues concerning the recognition conditions. If it then receives some clues about background noise (e.g., a radio), it might initiate a request to the user to establish a better acoustic environment. More specifically, the dialogue manager could generate the concepts to ask *Could you please turn off the radio?*

The quality of responses to users equally well can profit from information about the acoustic environment. To see this, imagine a situation where a police officer reports from the scene of an accident: In this situation, the acoustic conditions presumably are so adverse that recognition accuracy is inacceptable. Unlike in the scenario above, however, little can be done to change the environment. The appropriate action of the dialogue manager thus would be to make it clear to the officer that he is wasting his time in trying to get his message through.

Thus, with information on the acoustic conditions/environment, the dialogue in Figure 1 could become the one outlined in Figure 3.

[2] Take this as a metaphor. The architecture might in fact be black- or whiteboard-like.

user	Ich möchte gerne um zwei Uhr nach Hamburg fahren.
user	*I would like to travel to Hamburg at two o'clock.*
system	Sie sind wegen der Musik im Hintergrund leider schwer zu verstehen. Wäre es möglich, daß Sie nochmals anrufen, wenn Sie die Musik abgestellt haben?
system	*Unfortunately, I have got difficulties in understanding you due to the music in the background. Could you call again after having turned it off?*
user	Ja; bis gleich.
user	*Yes; until later.*

Figure 3: Dialogue utilizing acoustic clues

4 Techniques for Detecting Acoustic Clues

Work on the acoustic level that is suited for integration into a dialogue framework like that depicted above has not advanced far enough, yet. Only recently, in the background of the DARPA evaluations for speech recognition systems, the importance of the type of noise tracking that is needed has been realized. A component that not only detects but also classifies noise (as, e.g., music or street noise) has a good chance of becoming the first plug-and-play *spoken language interface entity (SLIE)*, and seems to be realizable by well-mastered techniques for speech recognition like Hidden Markov Models ((Rabiner, 1989)).

One approach for classifying acoustic conditions into different categories (e.g. background music) would be to use techniques like the ones used for non-word based topic spotting (see (Nowell and Moore, 1995)). The different categories of noise would correspond to topics, and typical sections of acoustic material from each category would correspond to keywords. Based on samples from each category/topic, keywords which are most useful in identifying this topic would then be extracted automatically. An incoming signal could then be classified as belonging to one of the categories, depending on which keywords appear most frequently.

Another approach is to build a simple Hidden Markov Model which gets trained for each category from the data in that category. An incoming signal could then be assigned to the category whose HMM gives the best match.

Research is also needed in the realm of dialogue management. It remains to be investigated in exactly which ways the acoustic information can be used. Obvious requests or follow-up questions like those examplified above are one option; more clever questions like *Our communication may proceed more smoothly if the system adapts to your acoustic conditions; shall this be done?* are another.

References

Martin Kay, Jean Mark Gawron, and Peter Norvig. 1994. *Verbmobil: A Translation System for Face-to-Face Dialog.* Number 33 in Lecture Notes. CSLI, Stanford, CA.

E. Maier, N. Reithinger, and Alexandersson J. 1997. Clarification dialogues as measures to increase robustness in a spoken dialogue system. In *Proceedings of the ACL/EACL Workshop on Spoken Dialog Systems*, Madrid.

Peter Nowell and Roger Moore. 1995. The application of dynamic programming techniques to non-word based topic spotting. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, Madrid.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257 – 286.